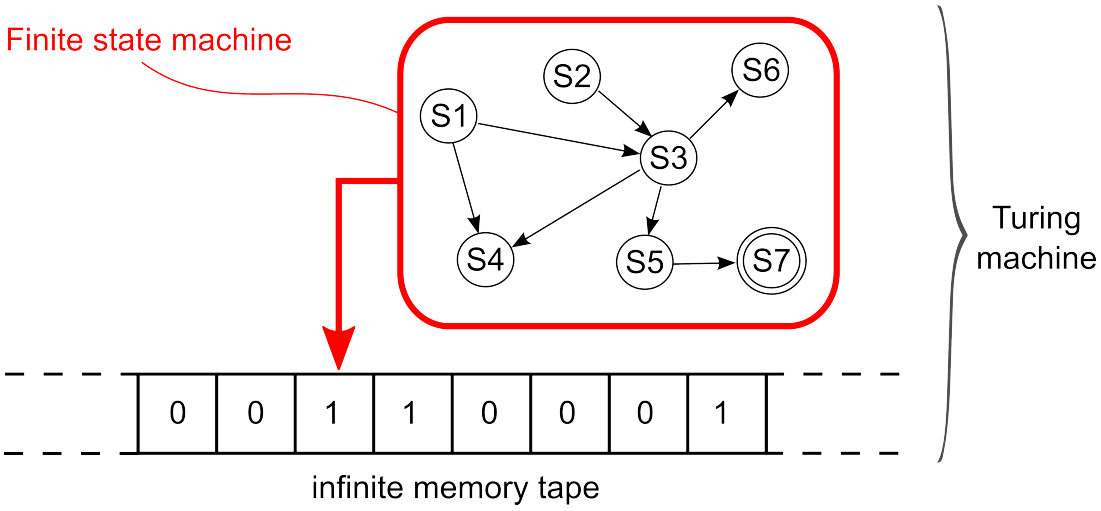
记忆的结构

甄景贤 (King-Yin Yan)

General.Intelligence@Gmail.com

Abstract. (Draft...) 智能系统中记忆的设计似乎是一个比较 open-ended question。

深度学习在近年很火，但 deep NN 的缺点是没有**记忆**。有限自动机 (finite state machines) 不是**全能**的计算器，它和 Turing machine 的分别就是缺少了那条「记忆磁带」：



(1)

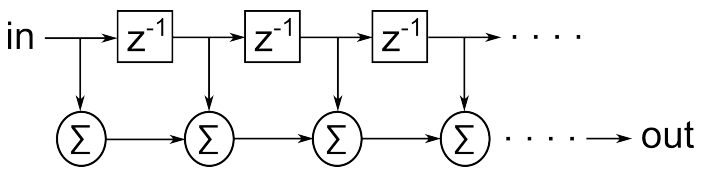
所以，深度神经网络 + 记忆 就可以变成 universal 的计算器。如何设计「可微分」的记忆是一个重要课题，因为可微分的记忆可以用 gradient descent 学习。

1 什么是 episodic memory?

在 [2] 中我们提出了智能系统的 minimal architecture，但这它没有对**历史**的记忆。换句话说，只能留意**当下**发生的事件，但不能记住一段**故事**。

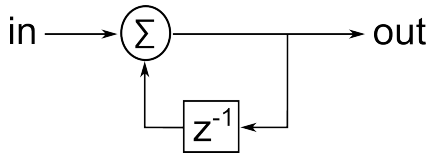
这牵涉到「什么是记忆?」的问题。在 minimal architecture 里，$\boldsymbol{F}$ 代表 "static knowledge"，亦即（相对地）**永恒不变**的知识 / 规律，而 $\boldsymbol{x}$ 代表当下的**状态** / 短期记忆，亦即 "dynamic knowledge"。Episodic memory 介乎「长期」与「短期」记忆之间。

在 **信息处理** (signal processing) 理论中，z-transform 以 z^{-1} 可以代表 **时间延迟** 1 步 (1-step time delay unit)。用一连串的 z^{-1} 可以组成长度为 k 单位的记忆：



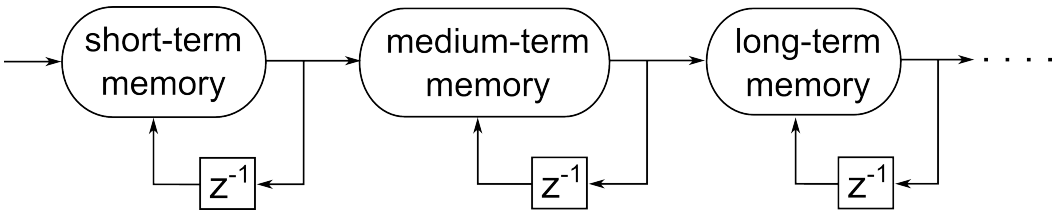
(2)

以上的无穷串列可以用这个 recursive 结构代替：



(3)

多层的 hierarchical 记忆或许可以这样实现：



(4)

Z-transform 的**连续时间**版本就是 Laplace transform。

上面的记忆结构来自 Simon Haykin 的经典著作《Signals and Systems》[1]。

2 Mental state $\boldsymbol{x}$ = working memory = 命题的集合

神经的状态空间由一些 thoughts（思维）组成，一个 thought 对应於逻辑中的一条**命题**，例如：

$$\boldsymbol{x} = \text{ 我正在上课} \wedge \text{ 我很肚饿} \wedge .... \tag{5}$$

这两个 thoughts 是独立的。也可以有另一个状态：

$$\boldsymbol{x}_2 = \text{ 我正在搭地铁} \wedge \text{ 我很肚饿} \wedge .... \tag{6}$$

Thoughts 独立的好处是表述的 economy（状态 $\boldsymbol{x}$ 分拆成若干独立的 thoughts）。$\boldsymbol{x}$ 是 M 个 thoughts 的集合，M 是 working memory 的大小。认知科学里有个说法 []：

$$\text{the size of human working memory } \approx 7 \pm 2 \text{ items} \tag{7}$$

但这些 items 可以有 chunking []，例如去超级市场买东西，「意大利粉、茄汁、芝士、香肠」这 4 件东西可以聚合成一个 chunk，这样可以记住的 items 数目多很多。

3 记忆 = 状态的历史

现在回看状态方程：

$$\boxed{\text{连续时间}} \qquad \dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) \tag{8}$$

$\dot{\boldsymbol{x}}$ 是状态 $\boldsymbol{x}$ 的改变方向，换句话说，这是描述**状态变化**的方程。但状态的变化 $\boldsymbol{f}(\boldsymbol{x})$ 不取决於状态的历史（$\boldsymbol{x} = \boldsymbol{x}|_t$ 仅代表状态在时间 t 的值），所以这个系统没有**记忆**。如果想要记忆的话，一个简单的做法是令：

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}|_t, \overbrace{\boldsymbol{x}(t \in \mathbb{T})}^{\text{状态 } \boldsymbol{x} \text{ 的历史}}) \tag{9}$$

$\boldsymbol{f}$ 就是我们的神经网络。$\boldsymbol{f}$ 的输入是：

- $\boldsymbol{x}|_t$ = 当下的状态 = 一些命题的集合；命题没有 time stamp
- $\boldsymbol{x}(t \in \mathbb{T})$ = 状态的历史；要考虑每个状态的 time stamp

4 联想记忆 = associative / content-addressable memory

如果用 content-addressable 的方法，似乎可以不用 time stamps，只要每个记忆 items 之间用「时间先后-link」连接就可以。人脑的记忆似乎是这样的。

5 结论

References

1. Simon Haykin and Barry van Veen. *Signals and systems*. John Wiley & Sons, 2003.
2. King Yin Yan, Juan Carlos Kuri Pinto, and Ben Goertzel. Wandering in the labyrinth of thinking – a cognitive architecture combining reinforcement learning and deep learning. (to be submitted AGI-2017).